Econ 551: Lecture Note 8
Increasing Efficiency of Maximum Likelihood
by Imposing Nonlinear Restrictions

Professor John Rust

1. Background: This note shows how we can increase the efficiency of maximum likelihood estimation (or any type of linear or nonlinear estimation for that matter) by imposing extra *a priori* information about the problem. The prior information about a problem comes in two forms:

1. The prior information embodied in the specification of the likelihood function, i.e. the parametric density $f(x|\theta)$ for the observable x's.

2. Additional prior information embodied in linear or nonlinear restrictions on the θ parameters. We can write these restrictions in the general form:

$$g(\theta) = 0 \tag{1}$$

where θ is a $K \times 1$ vector and g is a $J \times 1$ vector: i.e. we assume we have J linear or nonlinear restrictions on the K parameters of the model. In general J could be less than or equal to K.

2. Asymptotic Considerations Note that imposing the extra prior information $g(\theta) = 0$ is not necessary to obtain consistency of the maximum likelihood estimator. Assuming that the density $f(x|\theta)$ is correctly specified, the *information inequality* implies that the expected log-likelihood is maximized at $\theta = \theta^*$:

$$\theta^* = \underset{\theta \in \Theta}{argmax}\, E\{\log f(\tilde{x}|\theta)\} \equiv L(\theta). \tag{2}$$

If the additional restrictions $g(\theta) = 0$ are also correctly specified, i.e. $g(\theta^*) = 0$, then it is easy to see that asymptotically the constraint that $g(\theta) = 0$ is non-binding:

$$\theta^* = \underset{\theta \in \Theta}{argmax}\, L(\theta) \quad \text{subject to:} \quad g(\theta) = 0. \tag{3}$$

However in *finite samples* the restriction that $g(\theta) = 0$ will generally be binding, i.e. one can generally get a higher value for the log-likelihood of the *unrestricted maximum likelihood estimator* $\hat{\theta}_u$ defined by:

$$\hat{\theta}_u = \underset{\theta \in \Theta}{argmax}\, \frac{1}{N} \sum_{i=1}^{N} \log f(x_i|\theta) \equiv L_N(\theta), \tag{4}$$

than for the *restricted maximum likelihood estimator* $\hat{\theta}_r$ defined by:

$$\hat{\theta}_r = \underset{\theta \in \Theta}{argmax}\, L_N(\theta) \quad \text{subject to} \quad g(\theta) = 0. \tag{5}$$

Lemma: *Under the standard regularity conditions for maximum likelihood (see e.g. White, 1982) and assuming the likelihood $f(x|\theta)$ is correctly specified, then we have:*

$$\sqrt{N}\left[\hat{\theta}_u - \theta^*\right] \underset{d}{\Longrightarrow} N(0, I^{-1}(\theta^*)) \tag{6}$$

$$\sqrt{N}\left[\hat{\theta}_r - \theta^*\right] \underset{d}{\Longrightarrow} N(0, \Sigma), \tag{7}$$

where $I(\theta^*)$ is the $K \times K$ information matrix given by:

$$I(\theta^*) = E\left\{\frac{\partial \log f}{\partial \theta}(\tilde{x}|\theta^*)\frac{\partial \log f}{\partial \theta'}(\tilde{x}|\theta^*)\right\}, \tag{8}$$

Σ is the $K \times K$ covariance matrix given by:

$$\Sigma = I^{-1}(\theta^*) - I^{-1}(\theta^*)\nabla g(\theta^*)'\left[\nabla g(\theta^*)I^{-1}(\theta^*)\nabla g(\theta^*)'\right]^{-1}\nabla g(\theta^*)I^{-1}(\theta^*), \tag{9}$$

and $\nabla g(\theta^*)$ is the $J \times K$ matrix of partial derivatives of $g(\theta)$ with respect to θ:

$$\nabla g(\theta^*) = \frac{\partial g}{\partial \theta}(\theta^*). \tag{10}$$

Proof: We have already derived the asymptotic distribution of the unrestricted ML estimator in Econ 551, so we restrict attention to the asymptotic distribution of the restricted ML estimator $\hat{\theta}_r$. Note that we can convert the restricted ML problem in equation (5) to an unrestricted optimization problem by introducing a $J \times 1$ vector of *Lagrange multipliers* λ and forming the *Lagrangian function:*

$$\mathcal{L}(\theta, \lambda) = \underset{\theta \in \Theta}{argmax}\ \underset{\lambda \in R^J}{argmin}\ \frac{1}{N}L_N(\theta) + \lambda'\left[0 - g(\theta)\right]. \tag{11}$$

The first order or *Kuhn-Tucker* conditions for this constrained optimization problem are given by:

$$\nabla L_N(\hat{\theta}_r) - \hat{\lambda}_N'\nabla g(\hat{\theta}_r) = 0 \tag{12}$$

$$g(\hat{\theta}_r) = 0. \tag{13}$$

As is standard practice for determining the asymptotic distribution of nonlinear estimators, we "linearize" the problem by doing a first-order Taylor-series expansion of the Kuhn-Tucker conditions about the limiting values of $(\hat{\theta}_r, \hat{\lambda}_N)$ as $N \to \infty$. Since we assume the density $f(x|\theta)$ is the true data generating process when $\theta = \theta^*$, then under standard regularity conditions we can establish the uniform consistency of $L_N(\theta)$ to $L(\theta)$ which implies that $\hat{\theta}_r \to \theta^*$ with probability 1. If we let λ^* denote the Lagrange multiplier for the limiting version of equatio (11) when $N \to \infty$, we can see from the discussion of equations (2) and (3) that the constraint $g(\theta) = 0$ is non-binding asymptotically, which implies that $\lambda^* = 0$. So doing a Taylor expansion of equations (12) and (13) about $(\theta^*, 0)$ we get:

$$0 = \nabla L_N(\theta^*) - \nabla g(\overline{\theta})'(\hat{\lambda}_N - 0) + \left[-\hat{\lambda}_N'\frac{\partial^2 g}{\partial\theta\partial\theta'}(\overline{\theta}) + \frac{\partial^2 L_N}{\partial\theta\partial\theta'}(\overline{\theta})\right][\hat{\theta}_r - \theta^*] \tag{14}$$

$$0 = \nabla g(\overline{\overline{\theta}})[\hat{\theta}_r - \theta^*], \tag{15}$$

where $\overline{\theta}$ and $\overline{\overline{\theta}}$ are points on the line segment joining $\hat{\theta}_r$ and θ^*. We can write these equations in matrix form as:

$$\begin{bmatrix} \hat{\lambda}_N'\frac{\partial^2 g}{\partial\theta\partial\theta'}(\overline{\theta}) - \frac{\partial^2 L_N}{\partial\theta\partial\theta'}(\overline{\theta}) & \nabla g(\overline{\theta})' \\ \nabla g(\overline{\overline{\theta}}) & 0 \end{bmatrix}\begin{bmatrix} \hat{\theta}_r - \theta^* \\ \hat{\lambda}_N - 0 \end{bmatrix} = \begin{bmatrix} \nabla L_N(\theta^*) \\ 0 \end{bmatrix}. \tag{16}$$

Assuming the first matrix in equation (16) is invertible, we can solve to get:

$$\sqrt{N}\left[\begin{array}{c}\hat{\theta}_r-\theta^*\\ \hat{\lambda}_N-0\end{array}\right]=\left[\begin{array}{cc}\hat{\lambda}_N'\frac{\partial^2 g}{\partial\theta\partial\theta'}(\overline{\theta})-\frac{\partial^2 L_N}{\partial\theta\partial\theta'}(\overline{\theta}) & \nabla g(\overline{\theta})'\\ \nabla g(\overline{\overline{\theta}}) & 0\end{array}\right]^{-1}\left[\begin{array}{c}\sqrt{N}\nabla L_N(\theta^*)\\ 0\end{array}\right]. \tag{17}$$

The Central Limit Theorem implies that $\sqrt{N}\nabla L_N(\theta^*)\underset{d}{\Longrightarrow}N(0,I(\theta^*))$. The uniform law of large numbers implies that with probability 1 we have:

$$\left[\begin{array}{cc}\hat{\lambda}_N'\frac{\partial^2 g}{\partial\theta\partial\theta'}(\overline{\theta})-\frac{\partial^2 L_N}{\partial\theta\partial\theta'}(\overline{\theta}) & \nabla g(\overline{\theta})'\\ \nabla g(\overline{\overline{\theta}}) & 0\end{array}\right]\to\left[\begin{array}{cc}I(\theta^*) & \nabla g(\theta^*)'\\ \nabla g(\theta^*) & 0\end{array}\right]. \tag{18}$$

Furthermore it is straightforward to verify that the inverse of the matrix on the right hand side of (18) is given by

$$\left[\begin{array}{cc}I(\theta^*) & \nabla g(\theta^*)'\\ \nabla g(\theta^*) & 0\end{array}\right]^{-1}=\left[\begin{array}{cc}\Sigma & \Gamma'\\ \Gamma & \Omega\end{array}\right], \tag{19}$$

where Σ is given in equation (9), and Γ and Ω are given by:

$$\Omega = -[\nabla g(\theta^*)I^{-1}(\theta^*)\nabla g(\theta^*)']^{-1} \tag{20}$$

$$\Gamma = -\Omega\nabla g(\theta^*)I^{-1}(\theta^*). \tag{21}$$

Using this formula, it is then straigthforward to verify that $\sqrt{N}(\hat{\theta}_r-\theta^*)\underset{d}{\Longrightarrow}N(0,\Sigma)$ as claimed in equation (7). It is obvious from equation (9) that $\Sigma\leq I^{-1}(\theta^*)$, i.e. the addition of extra *a priori* information reduces the asymptotic covariance matrix of $\hat{\theta}_r$. This is just another way of saying that the restricted maximum likelihood estimator $\hat{\theta}_r$ is more efficient than the unrestricted maximum likelihood esitmator $\hat{\theta}_u$.

Exercise: What is the asymptotic distribution of $\sqrt{N}(\hat{\lambda}_N-0)$?